

Smoothing-Based Tests of Function Fit

Jeff Hart

Department of Statistics

Texas A&M University

Outline

- I.** Testing the fit of linear models

- II.** Testing function fit when a likelihood is specified

- III.** Examples

- IV.** Concluding remarks

For details, see

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, New York.

A Regression Model

We observe (\mathbf{x}_i, Y_i) , $i = 1, \dots, n$.

$$Y_i = r(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

- $\epsilon_1, \dots, \epsilon_n$ are i.i.d. errors with mean 0 and variance σ^2
- $\mathbf{x}_1, \dots, \mathbf{x}_n$ are d -dimensional vectors of co-variate values.
- The \mathbf{x}_i 's could be fixed or random, but if random, we condition upon their observed values.

Problem of Interest

Want a test of fit for a parametric regression model that will be reasonably powerful against a wide variety of alternatives.

Suppose, for example, we want to test the fit of the classical linear model

$$Y = \beta X + \epsilon.$$

For simplicity, let's say we have just one covariate, and want to test

$$H_0 : r(x) = \beta_0 + \beta_1 x \quad \text{for all } x.$$

Classical Tests of Fit

- I.** F -test that compares \hat{Y} with \bar{Y} : requires replicates and is often less powerful than tests based on smoothers

- II.** Reduction method: poor power against a number of reasonable alternatives

- III.** Likelihood ratio test: parametric, and so has same problem as **II.**

Smoothing-Based Lack-of-Fit Tests

Want to test

$$H_0 : r(x) = \beta_0 + \beta_1 x \quad \text{for all } x.$$

For a vector $\mathbf{a} = (a_1, \dots, a_n)'$, let $\hat{r}(x; \mathbf{a})$ be a linear, nonparametric smooth applied to \mathbf{a} , i.e.,

$$\hat{r}(x; \mathbf{a}) = \sum_{i=1}^n a_i w_i(x; h),$$

where h is a smoothing parameter. A nonparametric estimate of $r(x)$ is $\hat{r}(x; \mathbf{Y})$, which could be a kernel estimate, local polynomial, orthogonal series estimate,

Guiding principle behind smoothing-based tests

Regardless of whether or not H_0 is true, $\hat{r}(x; \mathbf{Y})$ will be relatively close to $r(x)$. So, consider how close $\hat{\beta}_0 + \hat{\beta}_1 x$ is to $\hat{r}(x; \mathbf{Y})$.

An obvious test statistic:

$$T = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \left[\hat{r}(x_i; \mathbf{Y}) - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

A better test statistic:

$$S = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \left[\hat{r}(x_i; \mathbf{Y}) - \hat{r}(x_i; \hat{\mathbf{Y}}) \right]^2,$$

where $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)'$ and

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n.$$

Can also write S in terms of residuals $e_i = Y_i - \hat{Y}_i$, $i = 1, \dots, n$:

$$S = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \hat{r}^2(x_i; \mathbf{e}).$$

S is preferable to T on the grounds that, under H_0 , bias is not a significant problem for S.

Practical Problems

- Need to approximate critical values for test based on S .
 - Large sample distribution: Hart (1997)
 - Gaussian-error approximations: Davies (1980)
 - Bootstrap

- Need to choose smoothing parameter
 - Data-driven method such as cross-validation: *Beware! This will affect sampling distribution of S .*
 - Choose h to maximize power: optimal h depends on the truth, so this is easier said than done. [See Kulasekera and Wang (1997).]
 - Compute P -value as a function of h : called *significance trace* by Young and Bowman (1995).

Multiple Regression

In principle, generalizing previous test to multiple regression is straightforward:

Apply a multivariate smoother to residuals from the hypothesized linear model, and proceed as before.

The only new problem that arises is the ever-present *curse of dimensionality*, which may lead to low power when d (the dimension of \mathbf{x}) is “large.”

We may recoup power against certain types of alternatives by using, for example, single index models or additive models.

A Likelihood-Based Method

When distributional properties of \mathbf{Y} are known or postulated, we may want to use a method based on the likelihood function. This is common in *generalized linear models*, as in binomial or Poisson regression.

Consider binomial regression, where we might want to test

$$H_0 : \text{logit}(r(x)) = \beta_0 + \beta_1 x.$$

A “nonparametric model” for $r(x)$ is orthogonal series of the form

$$\text{logit}(r(x)) = \beta_0 + \beta_1 x + \sum_{j=1}^{\infty} \beta_{j+1} u_j(x).$$

Test procedure of Aerts, Claeskens, and Hart (1999):

- Entertain “truncated” model

$$\beta_0 + \beta_1 x + \sum_{j=1}^m \beta_{j+1} u_j(x).$$

- Estimate $\beta_0, \beta_1, \dots, \beta_{m+1}$ by maximum likelihood.
- Choose m to maximize a modified AIC criterion:

$$AIC(m; C) = 2 \ell(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{m+1}) - Cm$$

- Reject H_0 iff maximizer of $AIC(m; C)$ is at least 1. Choose C to yield desired test level.

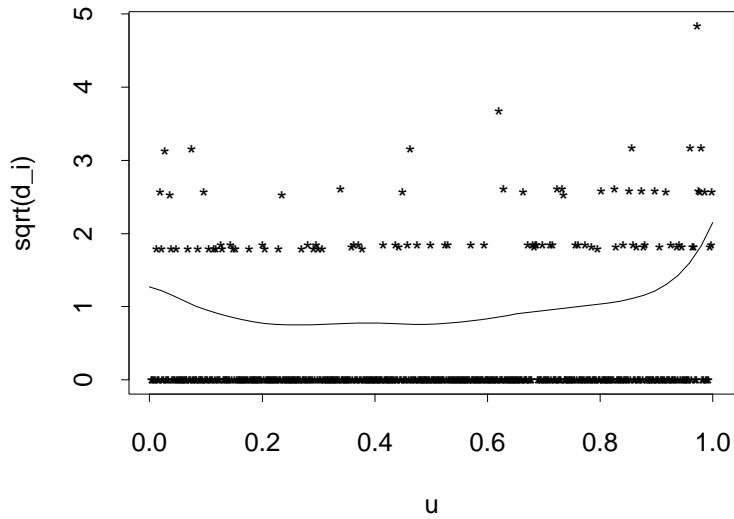
We'll call this an *order selection test*: Eubank and Hart (1992).

Examples

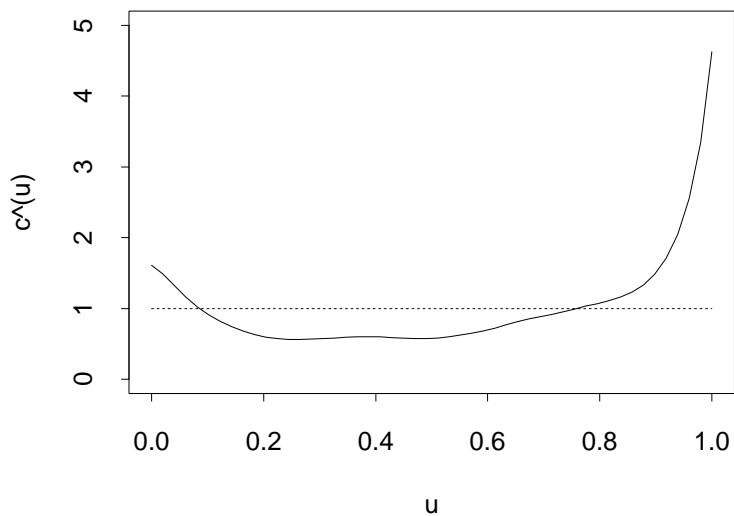
A study at Texas A&M Department of Horticulture involved cowpea plants. Test the hypothesis that a single gene is responsible for susceptibility of cowpeas to iron chlorosis.

Hypothesis of interest is equivalent to equality of two probability densities. Null hypothesis is true iff so-called *comparison density* is uniform.

- Empirical, or raw, comparison density has structure of regression data.
- Smooth raw comparison density using local polynomial.
- Using method described earlier, test hypothesis that comparison density is identical to 1.



*Square root of raw comparison density
and local quadratic estimate*



Estimated comparison density and uniform

The test statistic used was

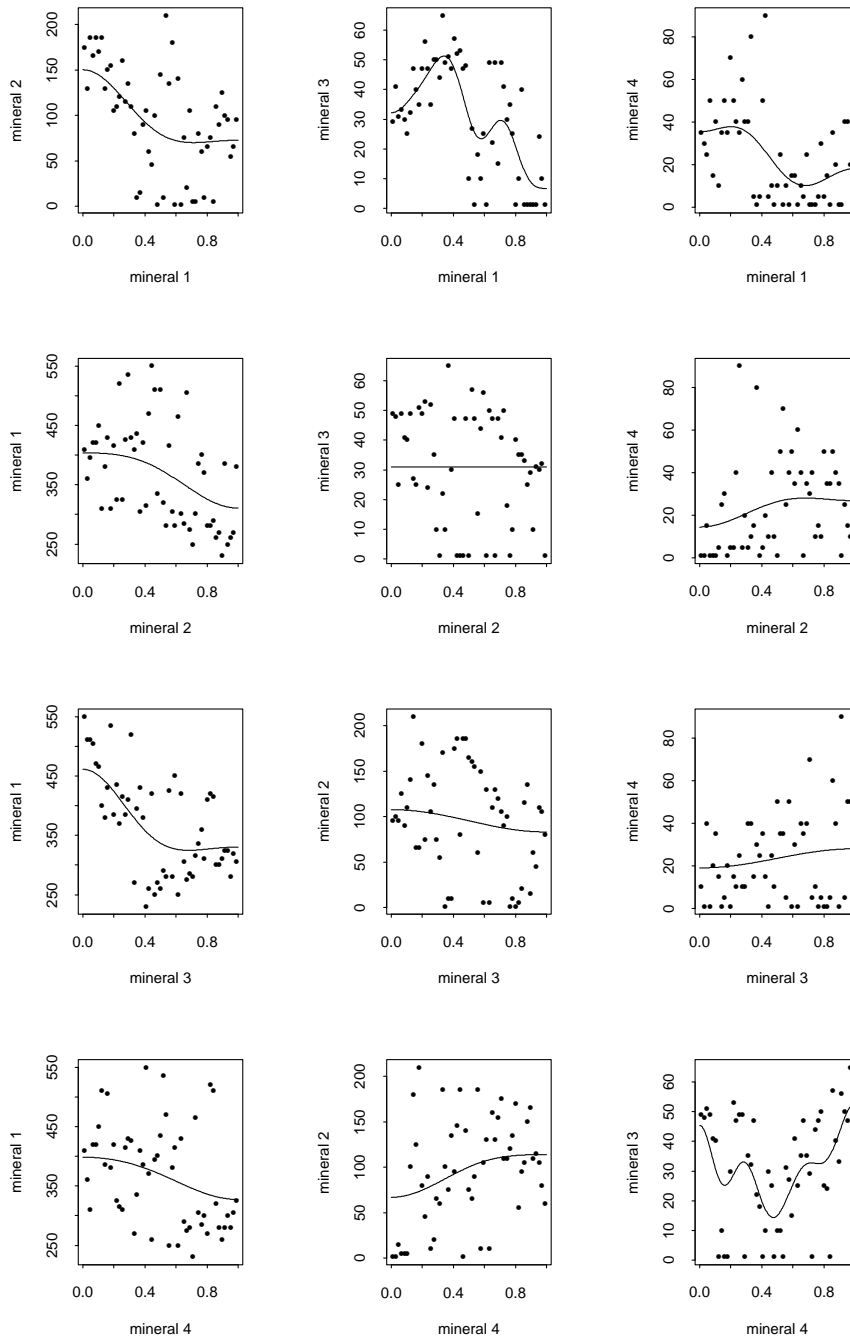
$$S = \frac{1}{n} \sum_{i=1}^n (\hat{c}(u_i) - 1)^2,$$

where \hat{c} is local linear estimate of comparison density with bandwidth chosen by *one-sided cross-validation* (Hart and Yi, 1998).

Bootstrap was used to approximate the distribution of S . Out of 1000 bootstrap samples, no bootstrap statistic was larger than observed value of S .

Estimated P -value is less than .001. Reject $H_0!$

Data from Chernoff (1973)



Level .05 order selection tests of association between any two variables in multivariate data set

- Lack-of-fit tests based on nonparametric smoothers are usually easy to apply and provide a good means of dealing with a-specific alternatives.
- I have some (rather piecemeal) *Splus* functions for doing smoothing-based lack-of-fit tests. For these, email me at

`hart@stat.tamu.edu`

- Upcoming software will be posted at

`http://stat.tamu.edu/~hart/`